IN THE SPECIFICATION

Please replace the paragraph beginning at page 2, line 19, with the following rewritten paragraph:

In general, some reasons why massive content collections are not well ~~utilised~~ utilized are:

- a user doesn't know that relevant content exists
- a user knows that relevant content exists but does not know where it can be located
- a user knows that content exists but does not know it is relevant
- a user knows that relevant content exists and how to find it, but finding the content takes a long time

Please replace the paragraph beginning at page 2, line 26, with the following rewritten paragraph:

The paper "Self ~~Organisation~~ Organization of a Massive Document Collection", Kohonen et al, IEEE Transactions on Neural Networks, Vol 11, No. 3, May 2000, pages 574-585 discloses a technique using so-called "~~self-organising~~ self-organizing maps" (SOMs). These make use of so-called unsupervised self-learning neural network algorithms in which "feature vectors" representing properties of each document are mapped onto nodes of a SOM.

Please replace the paragraph beginning at page 3, line 26, with the following rewritten paragraph:

The inventors of the present invention have ~~realised~~ realized that a user of such a Self ~~Organising~~ Organizing Map needs to have an awareness of the information associated with the selected node or nodes for efficient searching so that the user can check the information they have selected and/or to refine their search.

Please replace the paragraph beginning at page 4, line 29, with the following rewritten paragraph:

Figure 2 is a schematic flow chart showing the generation of a ~~self-organising~~ self-organizing map (SOM);

Please replace the paragraph beginning at page 6, line 28, with the following rewritten paragraph:

In a further example, the information items could be stored across a networked work group, such as a research team or a legal firm. A hybrid approach might involve some information items stored locally and/or some information items stored across a local area network and/or some information items stored across a wide area network. In this case, the system could be useful in locating similar work by others, for example in a large multi-national research and development ~~organisation~~ organization, similar research work would tend to be mapped to similar output nodes in the SOM (see below). Or, if a new television ~~programme~~ program is being planned, the present technique could be used to check for its originality by detecting previous ~~programmes~~ programs having similar content.

Please replace the paragraph beginning at page 7, line 16, with the following rewritten paragraph:

The process of generating a ~~self-organising~~ self-organizing map (SOM) representation of the information items will now be described with reference to Figures 2 to 6. Figure 2 is a schematic flow chart illustrating a so-called "feature extraction" process followed by an SOM mapping process.

Please replace the paragraph beginning at page 7, line 25, with the following rewritten paragraph:

The process of forming the ~~visualisation~~ visualization through creating feature vectors includes:

- Create "document database dictionary" of terms
- Create "term frequency histograms" for each individual document based on the "document database dictionary"
- Reduce the dimension of the "term frequency histogram" using random mapping
- Create a 2-dimensional ~~visualisation~~ visualization of the information space.

Please replace the paragraph beginning at page 8, line 1, with the following rewritten paragraph:

Considering these steps in more detail, each document (information item) 100 is opened in turn. At a step 110, all "stop words" are removed from the document. Stop-words are extremely common words on a pre-prepared list, such as "a", "the", "however", "about", "and", and "the". Because these words are extremely common they are likely, on average, to appear with similar frequency in all documents of a sufficient length. For this reason they serve little purpose in trying to ~~characterise~~ characterize the content of a particular document and should therefore be removed.

Please replace the paragraph beginning at page 8, line 15, with the following rewritten paragraph:

The result is a list of terms used in all the documents in the set, along with the frequency with which those terms occur. Words that occur with too high or too low a frequency are discounted, which is to say that they are removed from the dictionary and do not take part in the analysis which follows. Words with too low a frequency may be misspellings, made up, or not relevant to the domain represented by the document set. Words that occur with too high a frequency are less appropriate for distinguishing documents within the set. For example, the term "News" is used in about one third of all documents in a test set of broadcast-related documents, whereas the word "football" is used in only about 2% of documents in the test set. Therefore "football" can be assumed to be a better term for ~~characterising~~ characterizing the content of a document than "News". Conversely, the word "fottball" (a misspelling of "football") appears only once in the entire set of documents, and so is discarded for having too low an occurrence. Such words may be defined as those having a frequency of occurrence which is lower than two standard deviations less than the mean frequency of occurrence, or which is higher than two standard deviations above the mean frequency of occurrence.

Please replace the paragraph beginning at page 9, line 5, with the following rewritten paragraph:

It can be seen from this example how the histograms ~~characterise~~ characterize the content of the documents. By inspecting the examples it is seen that document 1 has more occurrences of the terms "MPEG" and "Video" than document 2, which itself has more occurrences of the term "MetaData". Many of the entries in the histogram are zero as the corresponding words are not present in the document.

Please replace the paragraph beginning at page 10, line 15, with the following rewritten paragraph:

It can be shown experimentally that by reducing a sparse vector from 50000 values to 200 values preserves their relative similarities. However, this mapping is not perfect, but suffices for the purposes of ~~characterising~~ characterizing the content of a document in a compact way.

Please replace the paragraph beginning at page 10, line 19, with the following rewritten paragraph:

Once feature vectors have been generated for the document collection, thus defining the collection's information space, they are projected into a two-dimensional SOM at a step 150 to create a semantic map. The following section explains the process of mapping to 2-D by clustering the feature vectors using a Kohonen ~~self-organising~~ self-organizing map. Reference is also made to Figure 5.

Please replace the paragraph beginning at page 10, line 24, with the following rewritten paragraph:

A Kohonen ~~Self-Organising~~ Self-Organizing map is used to cluster and ~~organise~~ organize the feature vectors that have been generated for each of the documents.

Please replace the paragraph beginning at page 10, line 26, with the following rewritten paragraph:

A ~~self-organising~~ self-organizing map consists of input nodes 170 and output nodes 180 in a two-dimensional array or grid of nodes illustrated as a two-dimensional plane 185. There are as many input nodes as there are values in the feature vectors being used to train the map. Each of the output nodes on the map is connected to the input nodes by weighted connections 190 (one weight per connection).

Please replace the paragraph beginning at page 11, line 4, with the following rewritten paragraph:

The closest node is designated the "winner" and the weights of this node are trained by slightly changing the values of the weights so that they move "closer" to the input vector. In addition to the winning node, the nodes in the ~~neighbourhood~~ neighborhood of the winning node are also trained, and moved slightly closer to the input vector.

Please replace the paragraph beginning at page 11, line 22, with the following rewritten paragraph:

A potential problem with the process described above is that two identical, or substantially identical, information items may be mapped to the same node in the array of nodes of the SOM. This does not cause a difficulty in the handling of the data, but does not help with the ~~visualisation~~ visualization of the data on a display screen (to be described below). In particular, when the data is ~~visualised~~ visualized on a display screen, it has been ~~recognised~~ recognized that it would be useful for multiple very similar items to be distinguishable over a single item at a particular node. Therefore, a "dither" component is added to the node position to which each information item is mapped. The dither component is a random addition of ±½ of the node separation. So, referring to Figure 6, an information item for which the mapping process selects an output node 200 has a dither component added so that it in fact may be mapped to any map position around a node 200 within the area 210 bounded by dotted lines on Figure 6.

Please replace the paragraph beginning at page 12, line 5, with the following rewritten paragraph:

At any time, a new information item can be added to the SOM by following the steps outlined above (i.e. steps 110 to 140) and then applying the resulting reduced feature vector to the "pre-trained" SOM models, that is to say, the set of SOM models which resulted from the ~~self-organising~~ self-organizing preparation of the map. So, for the newly added information item, the map is not generally "retrained"; instead steps 150 and 160 are used with all of the SOM models not being amended. To retrain the SOM every time a new information item is to be added is computationally expensive and is also somewhat unfriendly to the user, who might grow used to the relative positions of commonly accessed information items in the map.

Please replace the paragraph beginning at page 12, line 26, with the following rewritten paragraph:

Figure 7 schematically illustrates a display on the display screen 60 in which data sorted into an SOM is graphically illustrated for use in a searching operation. The display shows a search ~~enquiry~~ inquiry 250, a results list 260 and an SOM display area 270.

Please replace the paragraph beginning at page 12, line 29, with the following rewritten paragraph:

In operation, the user types a key word search ~~enquiry~~ inquiry into the ~~enquiry~~ inquiry area 250. The user then initiates the search, for example by pressing enter on the keyboard 70 or by using the mouse 80 to select a screen "button" to start the search. The key words in the search ~~enquiry~~ inquiry area 250 are then compared with the information items in the database using a standard keyword search technique. This generates a list of results, each of which is shown as a respective entry 280 in the list area 260. The display area 270 displays only points corresponding to each of the result items.

Please replace the paragraph beginning at page 13, line 4, with the following rewritten paragraph:

Because the sorting process used to generate the SOM representation tends to group mutually similar information items together in the SOM, the results for the search ~~enquiry~~ inquiry generally tend to fall in clusters such as a cluster 290. Here, it is noted that each point on the area 270 corresponds to the respective entry in the SOM associated with one of the results in the result list 260; and the positions at which the points are displayed within the area 270 correspond to the array positions of those nodes within the node array.

Please replace the paragraph beginning at page 14, line 25, with the following rewritten paragraph:

Floating RSVP in which the initial view of an image is at the ~~centre~~ center of a display area and small in size and which may be "out of focus". The image moves e.g. diagonally across the display area increasing in size. Many images are shown simultaneously.

Please replace the paragraph beginning at page 14, line 28, with the following rewritten paragraph:

Shelf RSVP in which successive images follow a predetermined trajectory (as if moving along a shelf for example) starting small at an emergence point at an edge of the display area and increasing to a maximum at the ~~centre~~ center of the display area and reducing again to disappear at another edge. Many images are displayed simultaneously moving along the trajectory.